**Course: Statistical and Predictive Modeling**

**Assignment 1: Distribution Analysis**

**Submitted by: Maisha Khatoon**

**Student ID: 100899259**

**Late Submission**

Contents

[Questions: 2](#_Toc159969746)

[1. Construction of a scatterplot with smoothing for mpg = miles per gallon vs wt. (Y-axis= mpg, X-axis= wt) 2](#_Toc159969747)

[2. Explanation of the following for the above scatterplot: 2](#_Toc159969748)

[a) Identifiable Trends (Direction, shape/form, strength, range, concentration of the data points, etc.) 2](#_Toc159969749)

[b) Anomalies (Any outliers, clusters, gaps, etc.) 3](#_Toc159969750)

[3. Construction of a side-by-side boxplot for disp. vs cyl. (Y-axis= disp, X-axis= cyl) 3](#_Toc159969751)

[4. Explanation of the following for the above boxplot: 3](#_Toc159969752)

[a) Measure of Central Tendency (Mean, Median, etc.) 3](#_Toc159969753)

[b) Measure of Dispersion (Range, Interquartile Range, Standard Deviation, etc.) 4](#_Toc159969754)

[R Scripts 4](#_Toc159969755)

[For the Scatterplot with Smoothing 4](#_Toc159969756)

[For the Boxplot construction 5](#_Toc159969757)

# Questions:

## 1. Construction of a scatterplot with smoothing for mpg = miles per gallon vs wt. (Y-axis= mpg, X-axis= wt)

A Scatterplot is a data visualization representation based on two variables, one variable along the Y-axis and the other along the X-axis. The plot comprises individual data points representing observations, arranged based on the values of the two variables under comparison. When analyzing data, scatterplots can be used to graphically detect patterns, trends, and correlations between the variables, such as whether there is a correlation or not. They are frequently used for exploratory data analysis and hypothesis testing in a variety of domains, including science, engineering, statistics, and the social sciences.

In this scatterplot, the weight is set to 1000lbs.

A graph of a line with black dots and a blue line

Description automatically generated

## 2. Explanation of the following for the above scatterplot:

### a) Identifiable Trends (Direction, shape/form, strength, range, concentration of the data points, etc.)

**Direction:**

X-axis= weight= wt= 1000lbs

Y-axis= Fuel efficiency in miles per gallon= mpg

**Shape/form:**

The above graph type is a linear association representation.

**Strength:**

Since most of the points on the graph follow the curve, it can be said that the graph strength is strong**.**

**Range:**

The formula for getting the range of the scatterplot,

The highest value along the Y-axis – The lowest value along the Y-axis

34-11

**Concentration of the data points:**

The data points are spread across the smoothed line in the graph.

### b) Anomalies (Any outliers, clusters, gaps, etc.)

## 3. Construction of a side-by-side boxplot for disp. vs cyl. (Y-axis= disp, X-axis= cyl)

A graph with different colored boxes

Description automatically generated

## 4. Explanation of the following for the above boxplot:

### a) Measure of Central Tendency (Mean, Median, etc.)

**Mean:**

The mean value for Cylinders across the three groups (4,6,8) is approximately = 260.

**Median:**

The median for the red plot = 110

The median for the green plot = 170

The median for the blue plot = 350

### b) Measure of Dispersion (Range, Interquartile Range, Standard Deviation, etc.)

**Range:**

Range of the red plot = 90 (140 - 50)

Range of the green plot = 100 (240 - 140)

Range of the blue plot = 200 (470 - 270)

**Interquartile Range:**

For the red plot,

Q3 – Q1 = 120 – 60 = 60

For the green plot,

Q3 – Q1 = 190 – 160 = 30

For the blue plot,

Q3 – Q1 = 390 – 300 = 90

**Standard Deviation:**

The values of the standard deviation for cylinders 4, 6, and 8 differ, reflecting the spread of the data; cylinder 8 has the highest value, and cylinder 6 has the lowest. These metrics are represented graphically by boxplots, which also shed light on how displacement data is distributed and varies between various cylinder groups.

It looks to be 18.75 for cylinder 4,

22.5 for cylinder 6,

and 50 for cylinder 8.

## R Scripts

### For the Scatterplot with Smoothing

# plot with both points and smoothed line

ggplot(carsDB, aes(x = wt, y = mpg)) + geom\_point() + geom\_smooth() + labs(title = "The Scatterplot with Smoothning line", x = "Weight (1000lbs)", y = "Fuel Efficiency (miles per gallon)")

### For the Boxplot construction

#BoxPlot

ggplot(data=carsDB, aes(factor(cyl), mpg))+ geom\_boxplot(aes(fill = factor(cyl))) + labs(title="side-by-side boxplot for disp. vs cyl")+ theme(plot.title = element\_text(hjust=0.5))